Zero-Resource Cross-Domain Named Entity Recognition

Zihan Liu, Genta Indra Winata, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
zihan.liu@connect.ust.hk

Abstract

Existing models for cross-domain named entity recognition (NER) rely on numerous unlabeled corpus or labeled NER training data in target domains. However, collecting data for low-resource target domains is not only expensive but also time-consuming. Hence, we propose a cross-domain NER model that does not use any external resources. We first introduce a *Multi-Task Learning* (MTL) by adding a new objective function to detect whether tokens are named entities or not. We then introduce a framework called *Mixture of Entity Experts* (MoEE) to improve the robustness for zero-resource domain adaptation. Finally, experimental results show that our model outperforms strong unsupervised cross-domain sequence labeling models, and the performance of our model is close to that of the state-of-the-art model which leverages extensive resources.

1 Introduction

Named entity recognition (NER) is a fundamental task in text understanding and information extraction. Recently, supervised learning approaches have shown their effectiveness in detecting named entities (Ma and Hovy, 2016; Chiu and Nichols, 2016; Winata et al., 2019). However, there is a vast performance drop for low-resource target domains when massive training data are absent. To solve this data scarcity issue, a straightforward idea is to utilize the NER knowledge learned from high-resource domains and then adapt it to low-resource domains, which is called cross-domain NER.

Due to the large variances in entity names across different domains, cross-domain NER has thus far been a challenging task. Most existing methods consider a supervised setting, leveraging labeled NER data for both the source and target domains (Yang et al., 2017; Lin and Lu, 2018).

However, labeled data in target domains is not always available. Unsupervised domain adaptation naturally arises as a possible way to circumvent the usage of labeled NER data in target domains. However, the only existing method, proposed by Jia et al. (2019), requires an external unlabeled data corpus in both the source and target domains to conduct the unsupervised cross-domain NER task, and such resources are difficult to obtain, especially for low-resource target domains. Therefore, we consider unsupervised zero-resource cross-domain adaptation for NER which only utilizes the NER training samples in a single source domain.

To meet the challenge of zero-resource cross-domain adaptation, we first propose to conduct multi-task learning (MTL) by adding an objective function to detect whether tokens are named entities or not. This objective function helps the model to learn general representations of named entities and to distinguish named entities from sequences in target domains. In addition, we observe that in many cases, different entity categories could have a similar or the same context. For example, in the sentence "Arafat subsequently cancelled a meeting between Israeli and PLO officials," the person entity "Arafat", can be replaced with an organization entity within the same context, which illustrates the confusion among different entity categories and makes zero-resource adaptation much more difficult. Intuitively, when the entity type of a token is hard to be predicted based on the token itself and the token's context, we want to borrow the opinions (i.e., representations) from different experts. Hence, we propose a Mixture of Entity Experts (MoEE) framework to tackle the confusion of entity categories, and the predictions are based on the tokens and the context, as well as all entity experts.

Experimental results show that our model is able to outperform current strong unsupervised cross-domain sequence tagging approaches, and reach comparable results to the state-of-the-art unsupervised method that utilizes extensive resources.

Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP-2020), pages 1–6

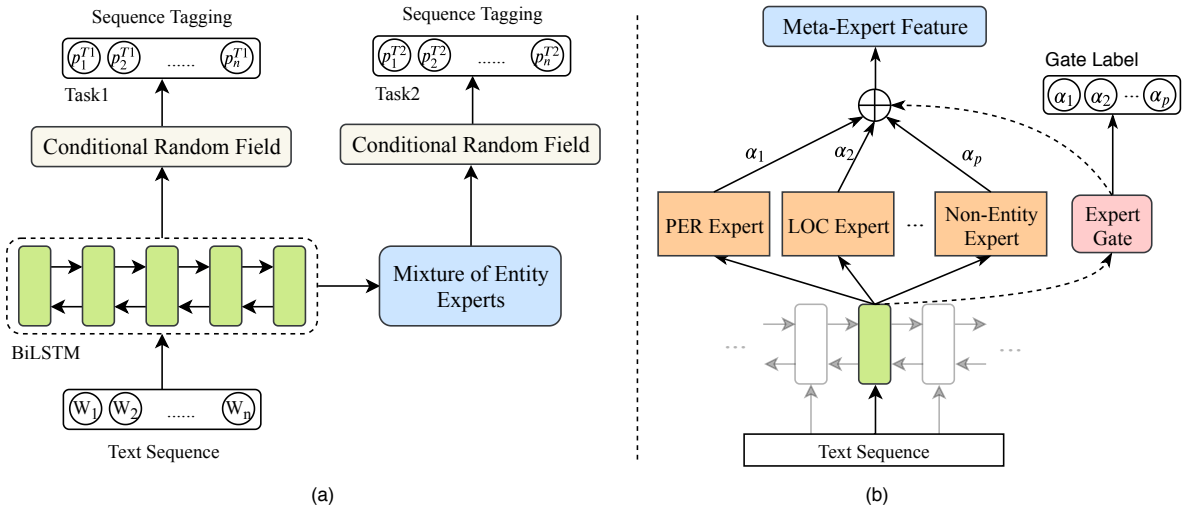

Figure 1: Model architecture (a) with multi-task learning and (b) with the Mixture of Entity Experts module.

2 Related Work

Most of the existing work on cross-domain NER has been to investigate the supervised setting, where both source and target domains have labeled data (Daume III, 2007; Obeidat et al., 2016; Yang et al., 2017; Lee et al., 2018). Yang et al. (2017) jointly trained models on the source and target domain with shared parameters. Lin and Lu (2018) added adaptation layers on top of existing models, and Wang et al. (2018) introduced label-aware feature representations for NER adaptation. Lee et al. (2018) utilized the idea of transfer learning by first initializing a target model with parameters learned from source-domain NER, and then using labeled target domain data to fine-tune the model. However, no prior work has focused on the unsupervised setting of cross-domain NER, except for Jia et al. (2019). In Jia et al. (2019), however, external unlabeled data corpora resources in both the source and target domains are required to train language models for domain adaptations. This limitation has motivated us to develop a model that doesn't need any external resources.

Tackling the low-resource scenario where there are zero or minimal existing resources has always been an interesting yet challenging task (Xie et al., 2018; Liu et al., 2019b; Lample et al., 2017; Conneau et al., 2017; Shah et al., 2019). Instead of utilizing large amounts of bilingual resources, Liu et al. (2019a,b) only utilized a few word pairs for zero-shot cross-lingual dialogue systems. Unsupervised machine translation approaches (Lample et al., 2017; Artetxe et al., 2017) have also been introduced to circumvents the need of parallel data. Winata et al. (2020) introduced the cross-accent speech recognition task and utilized meta-learning to cope with the data scarcity issue in target accents. Bapna et al. (2017) and Shah et al. (2019) proposed to do cross-domain slot filling with minimal resources. To the best of our knowledge, we are the first to propose methods on cross-domain adaptation for NER with zero external resources.

3 Methodology

As illustrated in Fig. 1, our model combines a bidirectional LSTM and conditional random field (CRF) into a BiLSTM-CRF structure (Lample et al., 2016) with MTL and MoEE modules. The parameters of BiLSTM are shared in the multi-task learning.

3.1 Multi-Task Learning

Due to the large variations of named entities across domains, unsupervised cross-domain NER models often suffer from an inability to recognize named entities. Hence, we propose to learn general representations of named entities and enhance the robustness for adaptation by adding an objective function to predict whether tokens are named entities or not, which is represented as Task_1 in Fig. 1(a). To do so, based on the original named entity labels for each token in the training set, we create another label set, which represents whether tokens are named entities or not. Specifically, in this process, all non-entity tokens are consistent with the original labels, and other tokens belonging to different entity categories are classified as being in the same class representing the general named entity. Task_2 in

Fig. 1(a) represents the original NER task, which is to predict a concrete category for each token. Let us denote $X = [w_1, w_2, ..., w_n]$ as the input text sequence, and the MTL can be formulated as:

$$[h_1, h_2, ..., h_n] = \text{BiLSTM}([w_1, w_2, ..., w_n]),$$
$$[p_1^{T_1}, p_2^{T_1}, ..., p_n^{T_1}] = \text{CRF}_1([h_1, h_2, ..., h_n]),$$
$$[m_1, m_2, ..., m_n] = \text{MoEE}([h_1, h_2, ..., h_n]),$$
$$[p_1^{T_2}, p_2^{T_2}, ..., p_n^{T_2}] = \text{CRF}_2([m_1, m_2, ..., m_n]),$$

where CRF_1 and CRF_2 denote the CRF layers for Task_1 and Task_2, respectively, and $[p_1^{T_1}, p_2^{T_1}, ..., p_n^{T_1}]$ and $[p_1^{T_2}, p_2^{T_2}, ..., p_n^{T_2}]$ represent the corresponding predictions.

3.2 Mixture of Entity Experts

Traditional NER models make predictions based on the features of the tokens and the context. Due to the confusion among different entity categories, NER models could easily overfit to the source domain entities and lose generalization ability to the target domain. Therefore, we introduce an MoEE framework, as depicted in Fig. 1(b). It combines representations generated by experts to produce the final prediction. In this way, the knowledge from different experts is incorporated to model the inherent confusion and improve the generalization ability to target domains.

Each entity category acts as an entity expert, which consists of a linear layer. Note that we consider the non-entity as a special entity category. The *expert gate* consists of a linear layer followed by a softmax layer, which generates the confidence distribution over entity experts. We use the gold labels in Task_2 to supervise the training of the expert gate. Finally, the meta-expert feature incorporates features from all experts based on the confidential scores from the expert gate. We formulate the MoEE module as follows:

$$[\text{expt}_i^1, \cdots, \text{expt}_i^E] = [\text{L}^1(h_i), \cdots, \text{L}^E(h_i)], \quad (1)$$
$$[\alpha_1, \cdots, \alpha_E] = \text{Softmax}(\text{Linear}(h_i)), \quad (2)$$
$$m_i = \sum_{a=1}^{E} \alpha_a * \text{expt}_i^a, \quad (3)$$

where m_i is the meta-expert feature for the i-th hidden state of the BiLSTM, where expt is the feature generated from the expert, and L denotes the linear layer. We show that the MoEE has E experts following the number of entity categories plus the non-entity category. The expert features are computed based on the BiLSTM hidden states, and the predictions are conditioned on the expert features and the hidden states, which makes cross-domain adaptation more robust.

3.3 Optimization

During training, we optimize for Task_1, Task_2 and the expert gate with cross-entropy losses $\mathcal{L}^{task1}$, $\mathcal{L}^{task2}$ and $\mathcal{L}^{gate}$, respectively, as we detail below:

$$\mathcal{L}^{task1} = \sum_{j=1}^{J} \sum_{k=1}^{|Y_j|} -\log(p_{jk}^{T_1} \cdot (y_{jk}^{T_1})^T), \quad (4)$$
$$\mathcal{L}^{task2} = \sum_{j=1}^{J} \sum_{k=1}^{|Y_j|} -\log(p_{jk}^{T_2} \cdot (y_{jk}^{T_2})^T), \quad (5)$$
$$\mathcal{L}^{gate} = \sum_{j=1}^{J} \sum_{k=1}^{|Y_j|} -\log(p_{jk}^{gate} \cdot (y_{jk}^{gate})^T), \quad (6)$$

where J and $|Y_j|$ denote the number of training data and the length of the tokens for each training sample, respectively; p_{jk} and y_{jk} denote the predictions and labels for each token, respectively; and the superscripts of p_{jk} and y_{jk} represent the tasks. Hence, the final objective function is to minimize the sum of all the aforementioned loss functions.

4 Experiments

4.1 Dataset

We take the CoNLL-2003 English NER data (Sang and De Meulder, 2003) containing 15.0K/3.5K/3.7K samples for the training/validation/test sets as our source domain. We take the dataset containing 2K sentences from SciTech News provided by Jia et al. (2019) as our target domain. The datasets in the source and target domains contain the same four types of entities, namely, PER (person), LOC (location), ORG (organization), and MISC (miscellaneous).

4.2 Experimental Setup

Embeddings We test our approaches on the FastText word embeddings (Bojanowski et al., 2017) and the pre-trained model BERT (Devlin et al., 2019). Entity names in the target domain are likely to be out-of-vocabulary (OOV) words because they don't usually exist in the source domain training set. FastText word embeddings are able to leverage the subword information and avoid the OOV problem, and BERT can solve this problem by using the BPE encoding. We try both freeze and unfreeze settings

Model	BERT	FastText	
	Fine-tune	unfreeze	freeze
Baseline			
Concept Tagger	67.14	62.34	66.86
Robust Sequence Tagger	67.31	63.66	68.12
Zero-Resource			
BiLSTM-CRF	67.55	63.18	68.21
w/ MTL	**68.76**	64.62	69.35
w/ MoEE	68.06	**65.24**	69.27
w/ MTL and MoEE	68.59	64.88	**69.53**
Using high-resource data in source and target domains			
Jia et al. (2019)	73.59		

Table 1: F1-scores on the target domain. Models are implemented based on the corresponding embeddings.

for FastText embeddings in the training. And for the BERT model, we add different modules (e.g., MoEE) on top to do fine-tuning.

Baselines Since we are the first to conduct zero-resource cross-domain NER, we compare our approach with strong unsupervised cross-domain sequence labeling models under minimal resources. **Concept Tagger** was proposed by Bapna et al. (2017) to utilize entity descriptions for unsupervised cross-domain utterance slot filling, and **Robust Sequence Tagger** (Shah et al., 2019) was introduced to combined both entity descriptions and a few examples from each entity category for the same unsupervised task. In addition, we also compare our approach with the following baselines **BiLSTM-CRF** (Lample et al., 2016), **BiLSTM-CRF w/ MTL**, and **BiLSTM-CRF w/ MoEE**, as well as with the state-of-the-art model of the unsupervised cross-domain NER from Jia et al. (2019) which utilizes a large corpus in both the source and target domains.

Training Details For FastText embeddings [1] based models, we use a BiLSTM with a 200-dimensional hidden state and two layers. The linear layer size for each entity expert is 200. An Adam optimizer with a learning rate of 1e-3, a batch size of 32, and a dropout rate of 0.3 are used to train our model. We utilize the binary models provided in FastText to obtain the embeddings for OOV words. For BERT-based models, given the strong textual understanding ability of the BERT model, we remove the BiLSTM from the text encoder, and only linear layer is utilized for sequence labeling (i.e., CRF layer is removed) (Devlin et al., 2019). As for the evaluation, we use the standard IOB (in-out-begin) format to calculate the F1-score.

[1] Available in https://fasttext.cc/docs/en/pretrained-vectors.html

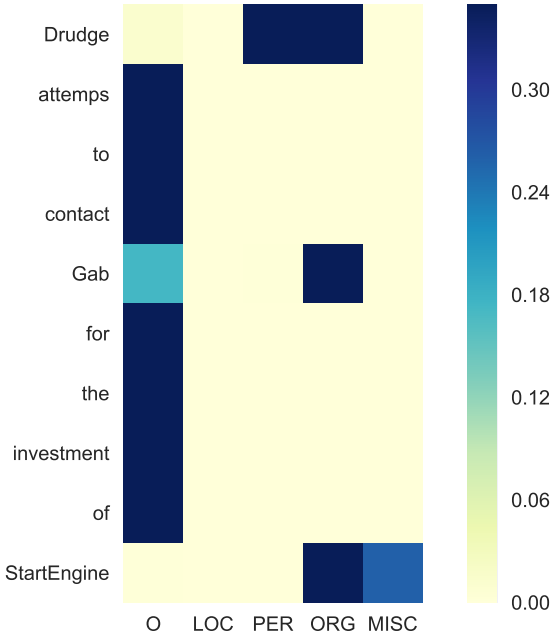

Figure 2: Confidence scores on different entity experts from the expert gate. "O" denotes non-entity expert.

4.3 Results & Discussion

From Table 1, our model combined with MTL and the MoEE outperforms the strong baselines Concept Tagger and Robust Sequence Tagger on all the embedding settings that we test. We conjecture that these two baselines, which utilize slot descriptions or slot examples, are suitable for limited slot names in the slot filling task, while they fail to cope with wide variances of entity names in the NER task across different domains, while our model is more robust to the domain variations. MTL helps our model recognize named entities in the target domain, while the MoEE adds information from different entity experts and helps our model detect the specific named entity types. Surprisingly, the performance of our best model (with freezed FastText embeddings) is close to that of the state-of-the-art model that needs a large data corpus in the source and target domains, which illustrates our model's generalization ability to the target domain.

We observe that the freezed FastText embeddings bring better performance than unfreezed ones. We conjecture that the embeddings could overfit to the source domain if we unfreeze them in the training. Additionally, using freezed FastText embeddings is slightly better than BERT fine-tuning.

We speculate that the reason is that NER is a word-level sequence tagging task, while the BERT model leverages subword embeddings, which could lose part of the word-level information for the task.

We visualize the confidence scores on different entity experts for each token in Fig. 2. The expert gate can align non-entity tokens to the non-entity expert with strong confidence. For some entity tokens, e.g., "Drudge", the expert gate gives high confidence on more than one expert (e.g., "PER" and "ORG") since the model is not sure whether "Drudge" is a "PER" or "ORG". Our model is expected to learn the "PER" and "ORG" expert representations based on the hidden state of "Drudge", which contains the information of this token and its context, and then combine the expert representations for the prediction.

5 Conclusion

In this paper, we propose a zero-resource cross-domain framework for the named entity recognition task, which consists of multi-task learning and Mixture of Entity Experts modules. The former learns the general representations of named entities to cope with the model's inability to recognize named entities, while the latter learns to combine the representations of different entity experts, which are based on the BiLSTM hidden states. Experimental results show that our model outperforms strong cross-domain sequence tagging models, and the performance is close to that of the state-of-the-art model that utilizes extensive resources.

Acknowledgments

This work is partially funded by ITF/319/16FP and MRP/055/18 of the Innovation Technology Commission, the Hong Kong SAR Government.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Ankur Bapna, Gokhan Tür, Dilek Hakkani-Tür, and Larry Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. *Proc. Interspeech 2017*, pages 2476–2480.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4(1):357–370.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019a. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. *arXiv preprint arXiv:1911.09273*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Rasha Obeidat, Xiaoli Z. Fern, and Prasad Tadepalli. 2016. Label embedding approach for transfer learning. In *ICBO/BioCreative*.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Darsh Shah, Raghav Gupta, Amir Fayazi, and Dilek Hakkani-Tur. 2019. Robust zero-shot cross-domain slot filling with example values. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5484–5490, Florence, Italy. Association for Computational Linguistics.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung. 2020. Learning fast adaptation on cross-accented speech recognition. *arXiv preprint arXiv:2003.01901*.

Genta Indra Winata, Zhaojiang Lin, Jamin Shin, Zihan Liu, and Pascale Fung. 2019. Hierarchical meta-embeddings for code-switching named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3532–3538.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *International Conference on Learning Representations*.